Multivariate Descriptive Statistics

Analysis of the dataset

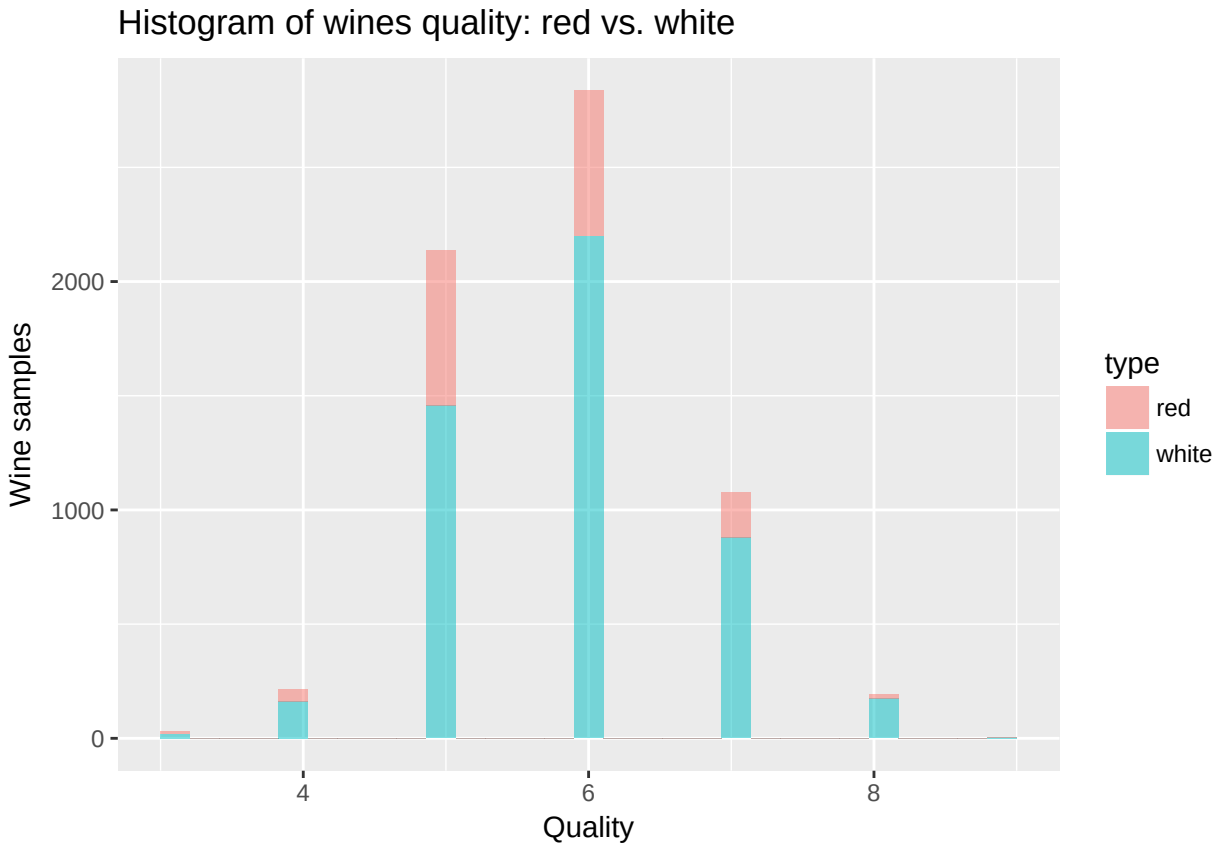
The dataset to analyze is composed of the red and white variants of the Portuguese "Vinho verde" wine. In total, there are 6495 different instances of wines, where 4898 are white wine and 1597 red wine. On the other hand, the dataset has 12 variables based on physicochemical tests on different wines plus two categorical variables, one for grading the wine quality by experts between 0 (very bad) and 10 (very excellent) and another binary variable, 0 = white wine and 1 = red wine.

The following Table 1 summarizes the distribution of the quantitatives variables, highlighting their minimum, maximum, mean and standard deviation values.

	min	**max**	**mean**	**SD**
fixed_acidity	3,800	15,900	7,215	1.296588
volatile_acidity	0,0800	1,5800	0,3396	0.164583
citric_acid	0,0	1,6600	0,3187	0.1452326
residual_sugar	0,600	65,800	5,444	4.758494
chlorides	0,00900	0,61100	0,05602	0.03503299
free_sulfur_dioxide	1,00	289,00	30,52	17.74849
total_sulfur_dioxide	6,0	440,0	115,8	56.52657
density	0,9871	1,0390	0,9947	0.002999095
pH	2,720	4,010	3,219	0.1608116
sulphates	0,2200	2,00	0,5313	0.148822
alcohol	8,00	14,90	10,49	1.192768

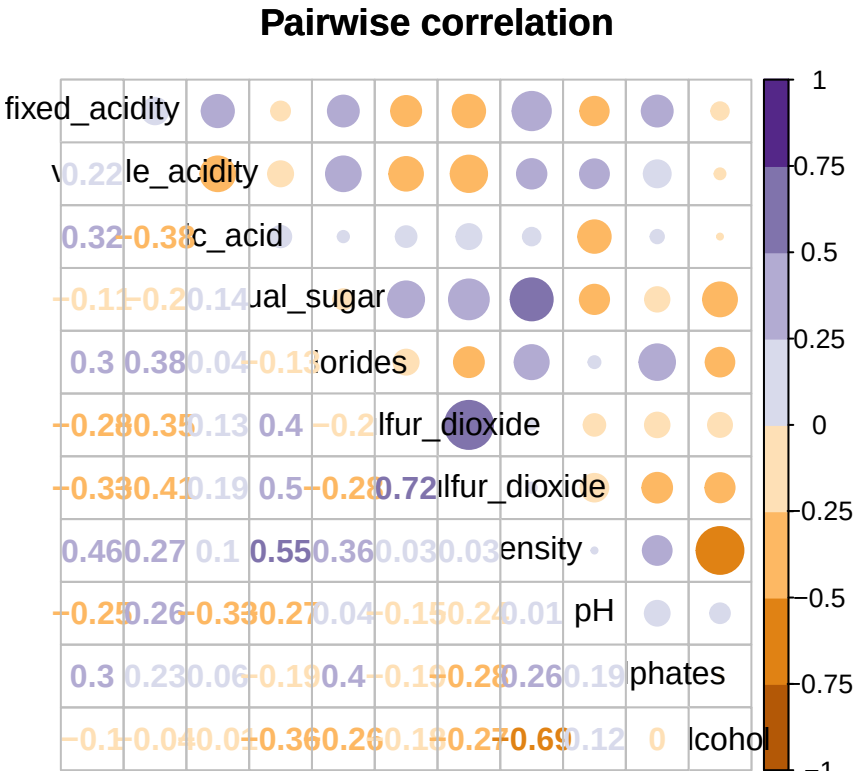
Table 1: Summary distribution of the quantitatives variables.

Quality histogram

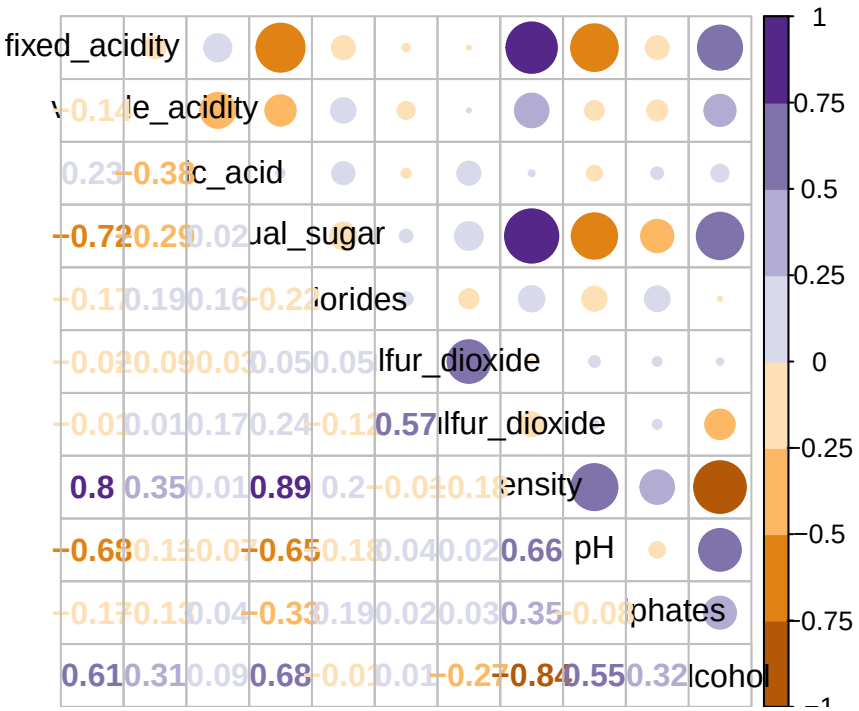


Correlation analysis

Let's now study the linear relationships between wine quantitative variables. The variables to be inspected are fixed_acidity, volatile_acidity, citic_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH, sulphates and alcohol.



As we can see, the linear relations between variables are weak. We can only assure that free_sulfur_dioxide obviously relates with the total_sulfur_dioxide. Also the residual sugar has effects on the density of the wines. To conclude, it is interesting to mention that the residual_sugar does not linearly relate with any of the other variables. Let's confirm our hypothesis using now the matrix of partial correlations:



Partial correlations, again, shows weak linear dependence between the variables. But fixed_acidity and

density, residual_sugar and density are indeed strongly correlated as we can see in the values.

Coefficients of determination

The coefficients of determination can also give us a good insight into linear relationships within the data.

	fixed_acidity	volatile_acidity	citic_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol
R-squared	0.7957831	0.4866891	0.3776351	0.8616690	0.3870994	0.5324288	0.6598145	0.9371628	0.6058479	0.3513432	0.7908480

Table 2: Coefficient of determination for each of the analyzed variables.

As the Table below displays, important linear relations are present in this set. Specifically, fixed_acidity, alcohol, density and residual_sugar are the most linearly explained by the other variables.

Effective dependence coefficient of the R matrix

Until this point, our analysis shows weak linear relations in this dataset. Further proof of this fact can be found in the effective dependence coefficient of the R matrix: $D(R) = 0.4052199$. This means that, altogether, linear dependences explain only 40% of the variability of the data.